



**BASIS**  
TECHNOLOGY

## GOVERNMENT USERS Conference

"Navigating the Human Terrain"  
College Park, MD, May 20-21, 2008



## A Linguistic Profile of the Persian Language and Dialects

Bushra Zawaydeh, Ph.D.  
*Senior Linguist*  
Basis Technology

# Arabic Script Languages

- Arabic, Persian, Urdu and Pashto have received considerable attention in the United States and globally, since the 9/11 attacks.
- The FBI has shifted its focus since then towards preventing future terrorist attacks, rather than solely investigating crime.



# National Security

- National Security depends on the ability to extract critical information from documents written in foreign languages. Information must be filtered and analyzed in time to act.



# Solution

- Basis Technology has tools that are able to extract meaningful intelligence from unstructured multilingual text.
- Basis Products handle: Middle Eastern, Asian, and European languages.
- This talk will focus on Persian.

# Basis Products Handling Persian

- *Rosette Base Linguistics*
- *Rosette Entity Extractor*
- *Rosette Name Translator*
- *Rosette Language Identification*
- *Digital Forensics*
- *Rosette Cross Language Toolkit*



Rosette Linguistics Platform

# Talk's Layout

- Part I: Background information about Persian.
- Part II: Complexities of the Perso-Arabic script & challenges for natural language processing processing.

# Persian Dialects

فارسی

| Country     | Known as      | Script       |
|-------------|---------------|--------------|
| Iran        | Farsi / Parsi | Perso-Arabic |
| Afghanistan | Dari          | Perso-Arabic |
| Tajikistan  | Tajik         | Cyrillic     |
| Uzbekistan  | Tajik         | Cyrillic     |

# Persian Language Map

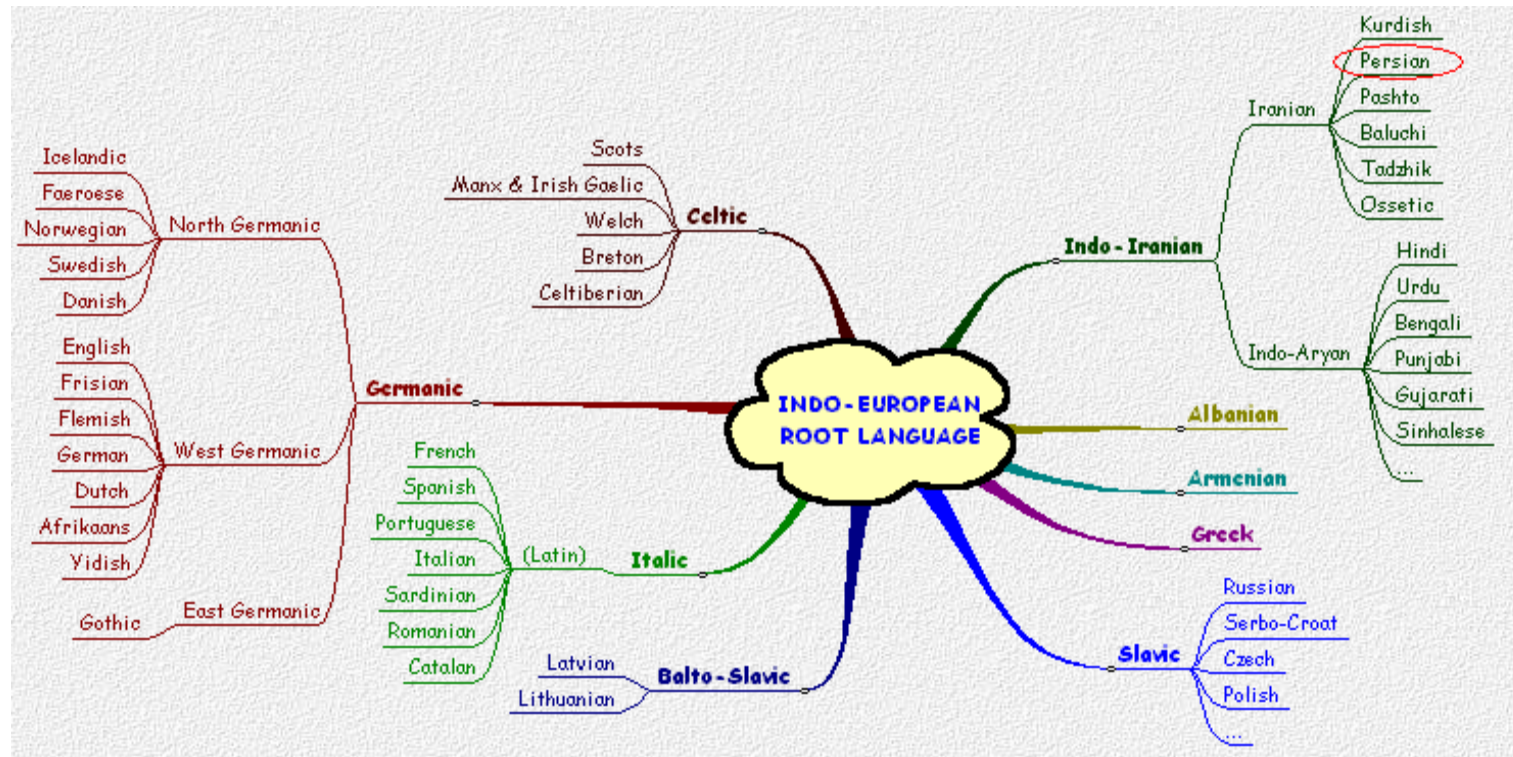


# Number of Persian Speakers

## CIA World Factbook:


- 64 million native speakers in Iran, Afghanistan, Tajikistan and Uzbekistan.
- About the same number of people in the rest of the world.

# Classification: Indo-European



# Persian, Farsi, or Parsi?



|         |  |
|---------|--|
| Parsi   | Original name in the native language   |
| Farsi   | <p>/p/ → [f]</p> <p>Name in Arabic and adopted in the native language</p>                                    |
| Persian | <p>Name in English</p>  |

# History of Persian

- Language of the Parsa people; the rulers of Iran between 550 -330 BC.
- Was the language of the Persian Empire, which spread between India, Russia, Persian Gulf.
- Empire spread 3,000 miles.
- Flourished for 200 years.

# Persian Empire



# Arabian Conquest

- 7<sup>th</sup> century AD.
- Prior to that the Persian religion practiced was Zoroastrianism.
- Shiite Islam became the national religion of Iran in the 16<sup>th</sup> century.



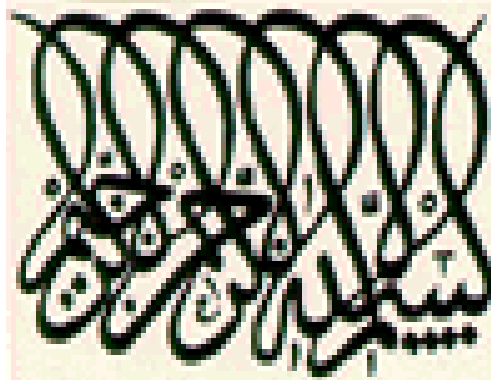
# Arabic Script

- After the Arabian conquest, Islam, Arabic, and the Arabic script were adopted.
- By the 11<sup>th</sup> century, Arabic was the common medium of expression from China to France.



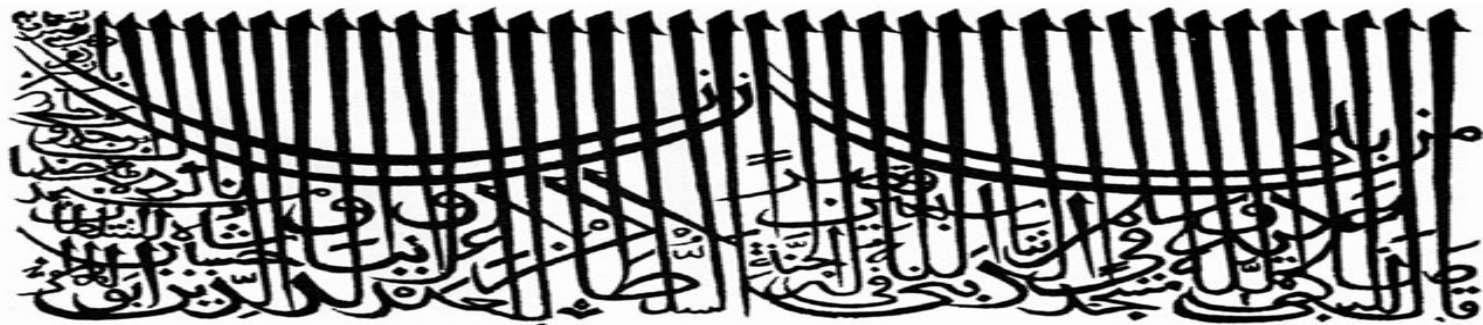
# Arabic

- Arabic is the language of Islam, since the Quran is written in Arabic.
- Until today, Arabic is used for prayer by Muslims.



# Arabic Script

- The Arabic script was adopted to write languages whose nations embraced Islam.
- Persian, Urdu, Pashtu, and even Turkish (until 50 years ago), used the Arabic script.



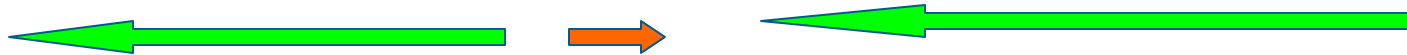
# Part II of This Talk

- Complexities of the Perso-Arabic Script:
  - Directionality
  - Letter features
  - Orthographic variations

# Complexities of Perso-Arabic Script

- Complex script - written right to left, but numbers are written left to right.

شمار فارسی‌دانان جهان را حدود ۱۱۰ میلیون نفر برآورد کرد



- The Unicode Standard's "Bidirectional Algorithm" provides specific mechanism to handle the storage and display of bi-directional text.

{ ن ك ب ه خ د ي ا ل } ه ت ب ج ز ن ي ك ل

- Arabic letters don't have one shape. Some join to adjacent letters.
- Some letters have up to 4 different shapes.

و و و و

ع ع ع ع



# Arabic Joining..

- The Unicode standard encodes characters not glyphs.
- There is one character for each Arabic letter.
- A render-time process selects the proper shape for each letter.

# Ligatures



- A ligature is a glyph that replaces two or more characters.
- The most common ligature is the “Lam-Aleph”.

Ⲛ = ⲛ = ⲗ + ⲁ

Ⲛ = ⲛ = ⲗ + ⲁ

# Example of Processing Arabic Script

|                                |               |         |
|--------------------------------|---------------|---------|
| Input text                     | Logical order | م ا ل س |
| After Bidirectional Algorithm  | Visual order  | س ل ا م |
| After Arabic Joining Algorithm | Glyph list    | س ل ا م |
| After Ligation                 | Glyph list    | س ل ا م |
| When Rendered                  | Output        | سلام    |

# Other Complexities

- Short vowels are not written.
- Capitalization does not exist.
  
- Such complexities make transliteration of names and named entity extraction more challenging.

# Persian Alphabet

|           |        |         |        |      |      |     |     |        |     |        |
|-----------|--------|---------|--------|------|------|-----|-----|--------|-----|--------|
| ذ         | د      | خ       | ح      | چ    | ج    | ث   | ت   | پ      | ب   | ا      |
| z         | d      | kh      | h      | ch   | j    | s   | t   | p      | b   | -      |
| [z]       | [d]    | [x]     | [h, Ø] | [tʃ] | [dʒ] | [s] | [t] | [p]    | [b] | [ʔ, ɔ] |
|           |        |         |        |      |      |     |     |        |     | [æ, Ø] |
| غ         | ع      | ظ       | ط      | ض    | ص    | ش   | س   | ژ      | ز   | ر      |
| gh        | '      | z       | t      | z    | s    | š   | s   | zh     | z   | r      |
| [ɣ]       | [ʔ, Ø] | [z]     | [t]    | [z]  | [s]  | [ʃ] | [s] | [ʒ]    | [z] | [r]    |
| [q, ɠ, x] |        |         |        |      |      |     |     |        |     |        |
| ی         | ه      | و       | ن      | م    | ل    | گ   | ک   | ق      | ف   |        |
| y         | h      | w       | n      | m    | l    | g   | k   | q      | f   |        |
| [j, i, e] | [h, Ø] | [v, u]  | [n]    | [m]  | [l]  | [g] | [k] | [q, ɠ] | [f] |        |
|           | [ɛ, æ] | [o, ow] |        |      |      |     |     |        |     |        |

# Persian Alphabet

- Added 4 more letters to the Arabic script.
- Extra letters:
  - U+067E *Peh* (پ), U+0686 *Tcheh* (چ),
  - U+0698 *Jeh* (ژ), U+06AF *Gaf* (گ)
- Modified letters:

| Character  | Isol | Fina | Medi | Init |
|--|------|------|------|------|
| U+064A <i>Arabic Letter Kaf</i><br>(Arabic Kaf)        | ك    | ك    | ك    | ك    |
| U+06CC <i>Arabic Letter Keheh</i><br>(Persian Kaf)     | ک    | ک    | ک    | ک    |
| U+064A <i>Arabic Letter Yeh</i><br>(Arabic Yeh)        | ي    | ي    | ي    | ي    |
| U+06CC <i>Arabic Letter Farsi Yeh</i><br>(Persian Yeh) | ی    | ی    | ی    | ی    |

# Next..

- Challenges of orthographic variations for NLP:
  - Persian Yeh and Keheh
  - Spelling ئ و أ
  - Spelling Arabic ة
  - Spelling Persian ه
  - Spelling affixes
  - Space
  - Arabic gutturals and interdentalals

# Persian *Yeh* an *Keheh*

- The first shipment of Microsoft Windows 2000 had the *Yeh* incorrectly mapped on the keyboard. Also, Times New Roman and Tahoma had wrong Persian *Yehs*.



# Persian *Yeh* and *Keheh*

- People may use the Arabic ones instead.
- “This has been further complicated by *helpful* software vendors in Iran selling fonts with the dots removed from Arabic Yeh!” (Esfahbod, 2004)
- More confusion → the Arabic Ya and Kaf are allowed to appear in Persian text, when Arabic is quoted.

# Example from RLP NE Extraction

- Named Entities extracted by RLP from a Persian text:

حسني مبارك رئيس جمهور مصر گفته است حماس و فتح جناحهای رقيب فلسطينی،

حسني مبارك رئيس جمهور مصر گفته است حماس و فتح جناحهای رقيب فلسطينی،

|              |   |
|--------------|---|
| GPE          | 2 |
| NATIONALITY  | 2 |
| ORGANIZATION | 4 |
| PERSON       | 2 |

# Persian ی و ا

- Some people omit the hamza completely. Hence these appear identical to اوی.
- The Persian yeh could be then underlyingly: a Persian yeh, ی, an aleph maqsura, or an Arabic yeh!

ي ی ی ی

# Variation in Spelling of Arabic ة

- There is no ة in Persian.
- If an Arabic word is borrowed, and has that letter, Persian:
  - Keeps it as is: ثقة الاسلام تبریزی
  - Turns it into a long "t": ثقت الاسلام تبریزی
  - Turns it into a "h". ثقه الاسلام تبریزی

# Persian ه

- Used to mark “ezafeh” constructions.
- Variations of how it is written:

1. Left unmarked as: ه

خانه من

---

2. “Correct spelling”: ه

خانهٔ من

---

3. Younger generation: هی

خانه‌ی من

# Variation in spelling affixes

- The Arabic prefix “bu al” / bol” is inconsistently spelled in Persian. Some names could have the prefix be spelled as “بل” and others could have it spelled as “بوال”:
  - بلهوس vs. بوالهوس (both are acceptable)
  - بلقاسم (incorrect) vs. بوالقاسم (correct)

# Space in Perso-Arabic

- Regular space.
  - Zero-Width Non-Joiner (ZWNJ)
  - Zero-Width Joiner (ZWJ)
- 
- This is a major challenge for tokenization of Persian. (Megerdooonian and Zajac, 2000)

# Zero-Width Non-Joiner

- Widely used in Persian.
- Used to write certain prefixes, suffixes, and compounds.
- Like the “hyphen” in English.
  - *hand writing*   *hand-writing*   *handwriting*

# Persian ZWNJ Example

- It is an invisible character. Placed between two characters that otherwise would be connected.
  - Correct form with ZWNJ: میخواهم
  - Incorrect (with space): می خواهم
  - Incorrect (joined) : میخواهم

All of these forms are attested.

# Zero-Width Joiner

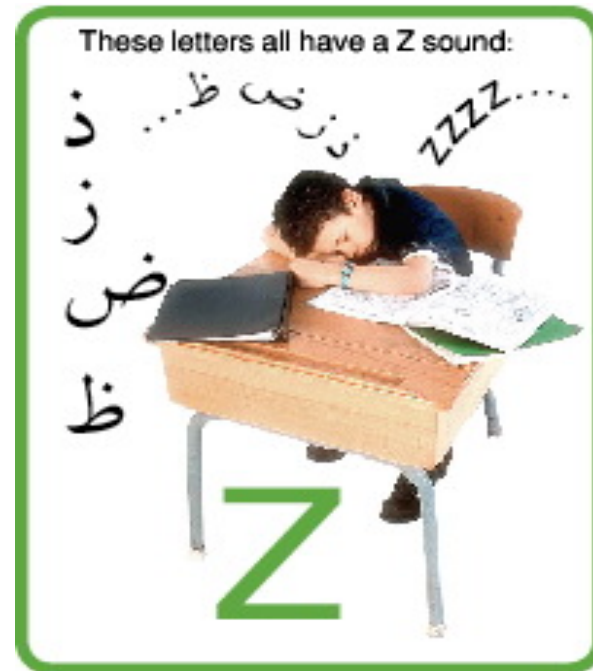
- Invisible character used to make a character that appears in isolation, to look like it is connected.

Example (Abbreviation of "Hejriye Shamsi"):

|                                |         |
|--------------------------------|---------|
| Incorrect "heh" (normal space) | ه . ش . |
| Correct "heh" (ZWJ space)      | ه . ش . |

# What does Persian do with Arabic Gutturals and Interdentals?

- Several letters are pronounced in one way:
  - [s]      ث، س، ص
  - [h]      ه، ح
  - [z]      ز، ذ، ض، ظ
  - [t]      ت، ط
  - [ʔ]      ع، ء
  - [q]      ق، غ



# How is it reflected in transliteration of Names?

- If there is no “th” in the Persian phonology, does it mean we won’t find “th” in Persian transliterated names?

|  |                      |
|--|----------------------|
| Engineer Abdul Lat <b>h</b> if<br>Roshan | انجینر عبدالطیف روشن |
| Ghaw <b>s</b> uddeen                     | غوث الدین            |

# /q/

- **ق** is pronounced “gh” between vowels.
  - aqa ‘man/sir’ → “aga” or “agha”.
  - Qermez ‘red’ → “Germez” not “ghermez”.

|                          |                     |
|--------------------------|---------------------|
| Doctor Nooragha Royen    | داکتر نور آقا رویین |
| Sayyad Khan Aqa Hussaini | سید خان آقا حسینی   |

- Data shows that people transliterate names according to:
  - Pronunciation
  - Spelling in the original script.
  - In a way to distinguish them from other sounds.

# Conclusion

- We presented an introduction to the Persian language, and linguistic challenges for NLP processing.



- Basis Technology has the tools, engineering, and linguistic knowledge that are capable of handling such challenges in our products.



مشکرم

Thank you

[bushraz@basistech.com](mailto:bushraz@basistech.com)